ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240643

Artwork Protection Against Neural Style Transfer Using Locally Adaptive Adversarial Color Attack

Zhongliang Guo[a,*], Junhao Dong[b], Yifei Qian[a], Kaixuan Wang[a], Weiye Li[a], Ziheng Guo[a], Yuheng Wang[a], Yanli Li[c], Ognjen Arandjelović[a] and Lei Fang[a]

[a]University of St Andrews
[b]Nanyang Technological University
[c]University of Sydney
ORCID (Zhongliang Guo): https://orcid.org/0000-0002-6025-3021

Abstract. Neural style transfer (NST) generates new images by combining the style of one image with the content of another. However, unauthorized NST can exploit artwork, raising concerns about artists' rights and motivating the development of proactive protection methods. We propose Locally Adaptive Adversarial Color Attack (LAACA), empowering artists to protect their artwork from unauthorized style transfer by processing before public release. By delving into the intricacies of human visual perception and the role of different frequency components, our method strategically introduces frequency-adaptive perturbations in the image. These perturbations significantly degrade the generation quality of NST while maintaining an acceptable level of visual change in the original image, ensuring that potential infringers are discouraged from using the protected artworks, because of its bad NST generation quality. Additionally, existing metrics often overlook the importance of color fidelity in evaluating color-mattered tasks, such as the quality of NST-generated images, which is crucial in the context of artistic works. To comprehensively assess the color-mattered tasks, we propose the Aesthetic Color Distance Metric (ACDM), designed to quantify the color difference of images pre- and post-manipulations. Experimental results confirm that attacking NST using LAACA results in visually inferior style transfer, and the ACDM can efficiently measure color-mattered tasks. By providing artists with a tool to safeguard their intellectual property, our work relieves the socio-technical challenges posed by the misuse of NST in the art community.

1 Introduction

Neural style transfer (NST) [8] is widely adopted in computer vision, where the distinctive stylistic elements of one image are algorithmically merged with the content features of another image using neural networks. While NST opens new avenues in artistic expression and digital image processing, it poses risks of misuse, particularly in the unauthorized use of curated artworks uploaded online. This concern has been raised by the British Broadcasting Corporation (BBC) [37], reporting that "many artists and photographers say they (a company named Stability AI) use their work without permission". Research efforts have been put into using the neural steganography techniques for post-violation accountability in post-NST images [7], but, to our knowledge, we witness an absence of proactive approaches that can protect artworks from unlawful replication and manipulation induced by NST before any financial and reputational damages occur.

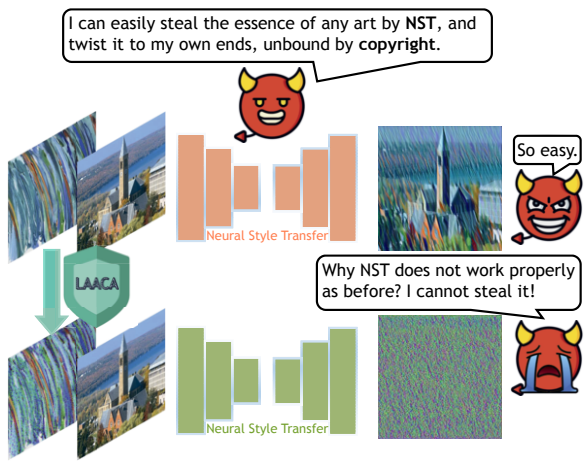

Figure 1. Role of our proposed method in preventing copyright infringement raised by unauthorized NST.

Adversarial attacks [36], a concept primarily explored in machine learning security, have shown promise in subtly altering input data to mislead neural networks. Several studies have demonstrated the effectiveness of adversarial attacks in various domains [11, 24, 43, 45, 47]. Inspired by the disruptive effect of adversarial attacks on machine learning-based systems, we propose to leverage this technique for artwork protection in the context of NST. By strategically embedding specific patterns or "adversarial perturbations" into digital artworks, we aim to systematically disrupt the unauthorized use of original artworks by AI models in advance. This approach offers a more robust and proactive defense mechanism compared to traditional methods like watermarking, as it directly targets the vulnerabilities of neural networks used in NST.

Color plays a crucial role in the perception and aesthetics of visual art [25, 40]. In the context of NST, color consistency is a fundamental aspect of style transfer algorithms [10, 21]. However, most existing reference-based image quality assessment metrics focus on image structure [39] or semantics [14, 32, 41], with limited attention given to color. This oversight leads to a lack of evaluation metrics

* Corresponding Author. Email: zg34@st-andrews.ac.uk

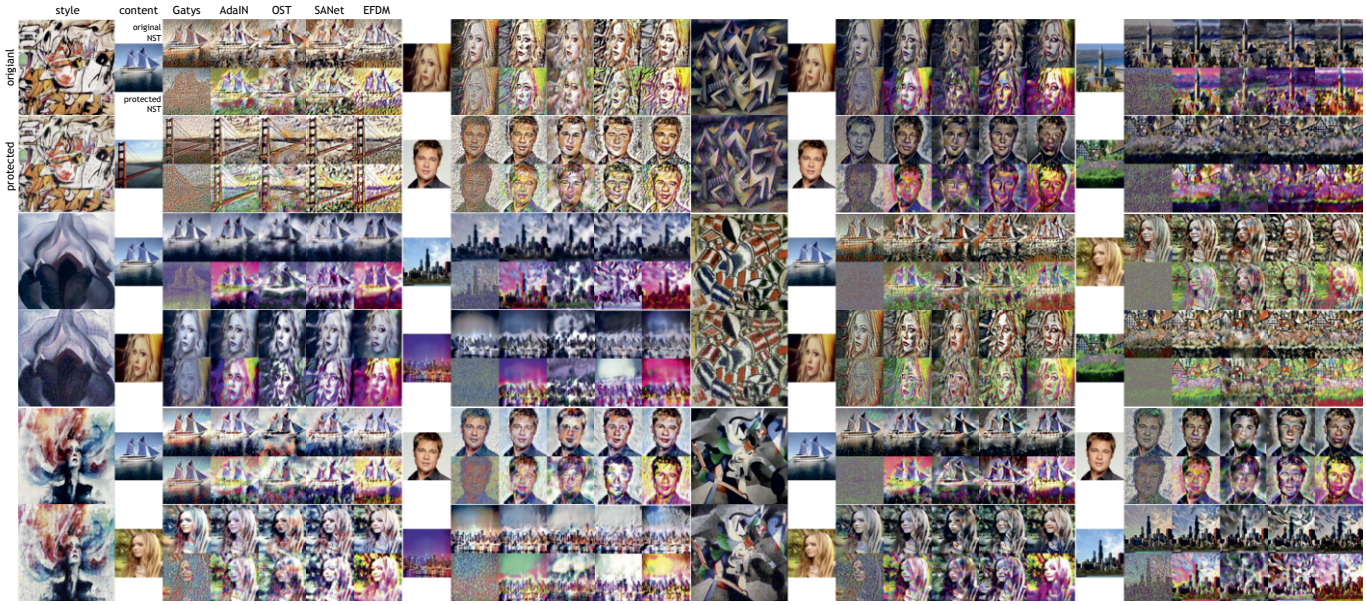

Figure 2. Adversarial examples against 5 NST methods on different style images and content images. For each item, on the far left, we exhibit the foundational images: the original style image at the top while the post-attack style image below it. Progressing to the right, the sequence is organized into four distinct groups for analysis. Each group commences with the content image, which provides the subject for the NST. Subsequent columns within each group depict the results of various NST methods (from left to right, they are Gatys [8], AdaIN [15], OST [26], SANet [31], and EFDM [42]). The top row across these groups showcases results from the original NST, the bottom row, in contrast, illustrates the post-attack NST outcomes. Specifically, most images displayed colors that were not visually present in the original style image and post-attack style image. Also, the textures in most images also suffer from disturbance.

specifically designed for color-sensitive tasks like NST. To address this issue, we propose the Aesthetic Color Distance Metric (ACDM), a novel metric that quantifies the color changes of images after undergoing certain transformations. By capturing color-related properties, ACDM provides a more comprehensive evaluation of color changes between pre- and post-manipulated images, which will also be helpful to exhaustively evaluate the proposed artwork protection method.

In light of the practical restrictions of artwork protection, we identify three main desiderata for image-level alterations tailored for NST: (a) acceptable perceptibility to the human eye that ensures the artwork's visual integrity, (b) effectiveness in disrupting the generation quality of most NST methods, and (c) a generic solution applicable to broad-spectrum of artworks. To address these requirements, we propose the Locally Adaptive Adversarial Color Attack (LAACA), a method that integrates adversarial techniques directly into the digital artwork creation process. LAACA leverages a frequency domain filter to divide the image into high-frequency and low-frequency content zones, and clips the perturbations in the high-frequency zone during each iteration of the attack. This approach ensures the visual integrity of the attacked images while effectively disrupting the color features and local texture details of the post-attacked NST images. Figure 2 demonstrates the impact of LAACA on NST outputs.

The main contributions of this work are as follows:

- We propose LAACA, a novel artwork protection method that proactively safeguards digital image copyrights by disrupting the NST generation through the addition of visually imperceptible perturbations to the input style image prior to the NST process.
- To address the limitations of existing image quality assessment metrics in evaluating color-mattered tasks, we introduce ACDM, a new metric that quantifies the color changes of images after undergoing certain transformations.

2 Related Works

In this section, we review the relevant literature in the fields of neural style transfer and adversarial attacks. We first discuss the evolution of NST algorithms, from the seminal work of Gatys et al. [8] to more recent advancements in Arbitrary Style Transfer. We then delve into the development of adversarial attacks, highlighting the shift towards a frequency domain perspective and the progress made in applying adversarial attacks to domains beyond image classification. Finally, we identify the research gap in adversarial attacks specifically targeting NST and position our work in the context of this gap.

Neural style transfer. Neural Style Transfer (NST) witnessed a foundational advancement with the work of Gatys et al. [8], which enabled the transfer of artistic style characteristics from one image to another through an iterative optimization process using the Gram Matrix. Building on this seminal work, subsequent research in NST explored alternatives to the Gram Matrix, offering improved stylization outcomes [9, 17, 27, 33]. A significant evolution in NST was the transition to non-iterable forms, known as Arbitrary Style Transfer (AST). A key development in this area was Adaptive Instance Normalization (AdaIN) [15], which simplified the style transfer process by training a decoder with fused statistical features of the style and content images. Furthermore, Lu et al. [26] offered a closed-form solution for NST, further streamlining the style transfer process. Park and Lee [31] integrated the attention mechanism into NST, enhancing the effectiveness of style transfer. Notably, Zhang et al. [42] updated the matching function in AdaIN by introducing Exact Feature Distribution Matching (EFDM), allowing for much better AST. It is important to highlight that our work does not aim to alter the parameters of NST algorithms; instead, we focus on manipulating the input style images to disrupt the style transfer process, offering a novel perspective on adversarial attacks in the context of NST.

Adversarial attack. The exploration of adversarial attacks against neural networks was pioneered by Szegedy et al. [36], who underscored the susceptibility of classification neural networks to perturbations in the input. Following this groundbreaking work, Goodfellow et al. [11] introduced one-shot adversarial perturbations by leveraging the gradients of neural networks to deceive classification models. Carlini and Wagner [1] proposed the first successful targeted attack on classification models trained with ImageNet [2]. Madry et al. [28] iteratively constrained image perturbations, allowing for more efficient convergence. Moosavi-Dezfooli et al. [30] proposed Universal Adversarial Perturbation, which can fool models with a single perturbation for arbitrary data. By introducing momentum in iterations, Dong et al. [5] further increased the transferability of adversarial samples.

A notable shift in the approach to adversarial attacks has been towards a frequency domain perspective, focusing on the role of frequency composition in the effectiveness and perceptibility of adversarial perturbations. Guo et al. [12] highlighted that solely using low-frequency noise can reduce computational costs for black-box attacks. Furthermore, Maiya et al. [29] offered that the frequency of noise in adversarial attacks is not strictly high or low but is related to the dataset. Advancing this inquiry, Jia et al. [16] explored generating perturbations in the frequency domain. Wang et al. [38] employed a conditional decoder to generate low-frequency perturbations, enabling a fast targeted attack. These developments suggest that considering adversarial attacks from a frequency domain standpoint could provide a more refined understanding and potentially enhance the effectiveness of attacks. Building upon these seminal advancements in adversarial attacks, the field has progressed into other domains of artificial intelligence [3, 4, 13, 44, 46].

To the best of our knowledge, there is only one method that attacks NST by disrupting *content images*, with no direct exploration of altering style images. The mentioned content-disruptive method, Feature Disruptive Attack (FDA) [6], manipulates the intermediate features of *content images* mapped by a neural network, resulting in distorted content in post-NST images while the applied style remains unchanged. However, the visual difference between pre- and post-attack images by FDA is slightly obvious. In contrast, our work focuses on adding imperceptible perturbations to *style images*, which results in significantly degraded post-NST images, regardless of the content images used. This content-independent method opens up new possibilities for adversarial attacks against unauthorized-NST usage, offers a more flexible and generalizable approach to disrupting NSTs.

3 Methodology

In this section, we first define the problem of artwork protection against NST in the adversarial attack framework via a simple yet unified formulation and propose our method. Additionally, we design a color-based metric named Aesthetic Color Distance Metric (ACDM) to assess the artistic style difference, which complements the existing Image Quality Assessment (IQA).

3.1 Problem Definition

We commence with a style image $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ from a set of style images $\mathcal{X}$, where C, H, and W denote the channels, height, and width of the image, respectively. Similarly, a content image $\boldsymbol{y} \in \mathbb{R}^{C\times H\times W}$ is selected from a content image set $\mathcal{Y}$.

The function $\boldsymbol{g} = \mathrm{NST}(\boldsymbol{x}, \boldsymbol{y})$ represents the neural style transfer process, which amalgamates the style of image $\boldsymbol{x}$ with the content of image $\boldsymbol{y}$. The output $\boldsymbol{g}$ denotes the resultant style-transferred image.

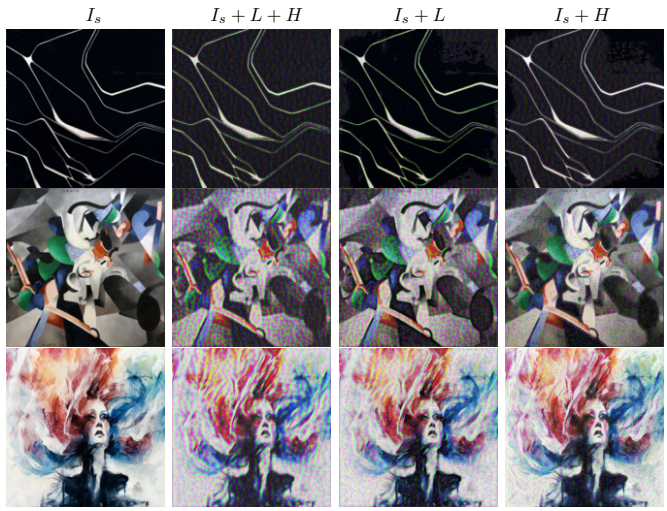

Figure 3. This figure illustrates various perturbation patterns applied to various style images, among perturbed images, "$I_s + H$" remains the best visual integrity. To be specific, I_s is the clean style image; L means low-frequency components of perturbations; H means high-frequency components of perturbations. Those perturbations are generated by our method with $k = 4$, $\alpha = 8$, $\epsilon = 80$, and $T = 100$. We separate the different frequency components by the kernel in Equation 4 with $k = 4$.

We introduce $\boldsymbol{x}^*$ as the protected style image generated from $\boldsymbol{x}$, where $\boldsymbol{x}^* = \boldsymbol{x} + \boldsymbol{\delta}$, and the difference vector $\boldsymbol{\delta} = \boldsymbol{x}^* - \boldsymbol{x}$ is the perturbation designed to disrupt NST. The essence of disrupting NST lies in creating a protected style image $\boldsymbol{x}^*$, visually similar to $\boldsymbol{x}$, yet significantly altering the NST generation combined with an arbitrary content image $\boldsymbol{y}$. To make changes visually imperceptible, the perturbation is restricted in an ℓ_p norm, denoted as $\|\boldsymbol{\delta}\|_p \leq \epsilon$, where ϵ is the defined budget of perturbations. Thus, the problem is:

$$\arg\max_{\boldsymbol{\delta}} \mathbb{E}_{\boldsymbol{y}\sim P_{\mathcal{Y}}}[D(\boldsymbol{g}, \boldsymbol{g}^*)] \quad s.t.\ \|\boldsymbol{\delta}\|_p \leq \epsilon\,, \tag{1}$$

we assume D is to measure the human perceptual distance; the post-protection NST output is $\boldsymbol{g}^* = \mathrm{NST}(\boldsymbol{x}^*, \boldsymbol{y})$; the expectation is taken with the content images' population distribution $P_{\mathcal{Y}}$.

Previous unauthorized NST prevention methods, like neural steganography or watermarking, offer post-violation accountability but rely on detecting infringement after the fact. Conversely, the adversarial attack method proactively introduces imperceptible perturbations that degrade NST output quality, deterring potential infringers by rendering the resulting images unsuitable for their intended purpose. Figure 1 of supplementary material[1] illustrates the difference of those different technique approaches.

3.2 Locally Adaptive Adversarial Color Attack

Building upon the Equation 1, there will be a challenge: considering the extensive variety of content images $\boldsymbol{y} \in P_{\mathcal{Y}}$ for NST, it becomes impractical to enumerate and process every content and style image combinations for each single style image. For example, using the ImageNet [2] as a benchmark, we recognize that a representative content image subset would include around 1,000 categories, and each category has a significant in-class variance. We assume that sampling at least 10 images from each class would slightly cover this in-class

[1] https://github.com/ZhongliangGuo/LAACA/blob/main/supp.pdf

variance, leading to a minimum of 10,000 samples for the subset. This approach poses a substantial computational challenge, particularly for iterative NST methods where time costs are significant.

To address this challenge, we try to make a paradigm shift to a method that is only reliant on style image $\boldsymbol{x}$. In this new framework, we introduce an amortized encoder, denoted as f, coupled with a surrogate loss function, J. The encoder f is designed to take an input style image $\boldsymbol{x}$ and output its content-less style representation. This representation is a distilled essence of the style image's characteristics, capturing features that define its unique artistic style. The surrogate loss function J then measures the disparity between two style representations generated by f. This approach allows us to approximate the $\mathbb{E}_{\boldsymbol{y}\sim P_{\mathcal{Y}}}[D(\mathrm{NST}(\boldsymbol{x},\boldsymbol{y}),\mathrm{NST}(\boldsymbol{x}^*,\boldsymbol{y}))]$ with $J(f(\boldsymbol{x}),f(\boldsymbol{x}^*))$, which is content-independent. Therefore, our task is to maximize the difference between the style representations of the original and protected style images, while ensuring the perturbation $\boldsymbol{\delta}$ is within a pre-defined budget limit ϵ, resulting acceptable visual integrity:

$$\arg\max_{\boldsymbol{\delta}} J(f(\boldsymbol{x}),f(\boldsymbol{x}^*)),\ s.t.\ \|\boldsymbol{\delta}\|_p \le \epsilon\,. \tag{2}$$

Inspired by previous works [20, 38], we restrict the perturbation $\boldsymbol{\delta}$ to be frequency-adaptive such that the visual effect is better preserved. This is motivated by the observation that the high correlation between different frequency components and visual effects. This finding guides us to embed adversarial perturbations within the high-frequency areas of the style image. This idea is demonstrated in Figure 3, where restricting perturbations to the high-frequency zone maintains higher visual integrity compared to other patterns. More formally, denote the pixel set of high-frequency components of an image $\boldsymbol{x}$ as $M(\boldsymbol{x}) \subseteq \{(i,j) \mid i=1,\dots,H, j=1,\dots,W\}$, and its complement $\neg M(\boldsymbol{x})$ becomes the corresponding low-frequency pixel set. Thus, the problem is formulated as:

$$\begin{aligned}&\arg\max_{\boldsymbol{\delta}} J(f(\boldsymbol{x}),f(\boldsymbol{x}^*)),\\ &s.t.\ \|\boldsymbol{\delta}\|_p \le \epsilon \text{ and } \boldsymbol{\delta}[i,j]=0, \text{ for } (i,j)\in\neg M(\boldsymbol{x}).\end{aligned} \tag{3}$$

Frequency separator. We employ a low-pass Gaussian filter to separate different frequency components from an image:

$$G_k(i,j) = \frac{1}{2\pi k^2} e^{-\frac{i^2+j^2}{2k^2}}, \tag{4}$$

where $G_k(i,j)$ denotes the value of the Gaussian kernel at position (i,j). The standard deviation of the kernel, σ, is determined by k. The kernel size is $(4k+1)\times(4k+1)$. The output vector from this kernel is the low-frequency components of the image. By using the above frequency separator, the high-frequency components of the style image $\boldsymbol{x}$ become:

$$\begin{aligned}M(\boldsymbol{x}) = \{(h,w) | \boldsymbol{x} - G_k(h,w) > 0;\\ h,w\in\mathbb{N}; 1<h\le H, 0<w\le W\},\end{aligned} \tag{5}$$

where $\mathbb{N}$ represents the set of natural number. The pixels in an image $\boldsymbol{x}$ belonging to $M(\boldsymbol{x})$ are denoted as high-frequency zone, and pixels belonging to $\neg M(\boldsymbol{x})$ is denoted as low-frequency zone.

Encoder. For the amortized encoder f, we utilize a pre-trained VGG [35], using its several layer outputs as feature extraction encoders. This choice is inspired by established arbitrary style transfer methods that effectively extract images' style representation from intermediate network layers. Each layer of the network is denoted as l, collectively denoted as a set L, $f^l(\boldsymbol{x})$ indicates the mapped result of intermediate layer l of the style image $\boldsymbol{x}$.

Algorithm 1 Locally Adaptive Adversarial Color Attack (LAACA)

Input: A style transfer encoder f with style loss function J; a real style image $\boldsymbol{x}$; a Gaussian kernel G_k with kernel size k
Parameter: The attack step size α ; ℓ_∞-norm perturbation radius ϵ; iterations T
Output: Attacked style image $\boldsymbol{x}^*$

$clamp_m^n$ restricts a value to be within the range $[m,n]$
randomly generate $\boldsymbol{\delta_0}$ in $[0,2]$ to avoid the gradient is 0 in loops
$\boldsymbol{x}_0^* = clamp_0^{255}\left[\boldsymbol{x} + M(\boldsymbol{\delta_0})\right]$
for $t=0$ to $T-1$ **do**
  $\boldsymbol{x}_t^*$ requires gradient
  Input $\boldsymbol{x}_t^*$ and $\boldsymbol{x}$ to f and obtain the gradient $\nabla_{\boldsymbol{x}} J\left(f\left(\boldsymbol{x}_t^*\right), f\left(\boldsymbol{x}\right)\right)$
  Update $\boldsymbol{x}_t^*$ by accumulating the signed gradient $\boldsymbol{x}_t^* = \boldsymbol{x}_t^* + \alpha \cdot sign\left[\nabla_{\boldsymbol{x}} J\left(f\left(\boldsymbol{x}_t^*\right), f\left(\boldsymbol{x}\right)\right)\right]$
  Get the perturbation and apply the mask on it $clamp_{-\epsilon}^{\epsilon}\boldsymbol{\delta_t}\left[\neg M(\boldsymbol{x})\right] = 0$
  Update $\boldsymbol{x}_{t+1}^*$ by the masked perturbation $\boldsymbol{x}_{t+1}^* = clamp_0^{255}\left[\boldsymbol{x} + \boldsymbol{\delta_t}\right]$
end for
return $\boldsymbol{x}_t^*$

Color disruptive loss function. As for the surrogate loss function J, our goal is to measure aspects of the neural network's intermediate layer mappings that represent color. This leads us to consider the mean μ and standard deviation σ, which are important in the neural network feature representation in terms of NST. Evidenced by Zhang et al. [42], who tested the influence of μ and σ, when only matching the μ of content representations with μ of style representations, the color of the post-NST image will be the same as that of its style counterpart; in contrast, when only matching σ, the texture will be similar. That is, μ represents the color, while the σ represents the contrast and texture variations, both of which significantly contribute to an image's style. Therefore, we design the surrogate loss function targeting those two statistics:

$$\begin{aligned}J\big(f(\boldsymbol{x}),f(\boldsymbol{x}^*)\big) = \sum_{l\in L}\Big(&\big(\mu(f^l(\boldsymbol{x})) - \mu(f^l(\boldsymbol{x}^*))\big)^2\\ &+\big(\sigma(f^l(\boldsymbol{x})) - \sigma(f^l(\boldsymbol{x}^*))\big)^2\Big),\end{aligned} \tag{6}$$

where μ is a function to get the mean of feature in each channel, and σ is a function to get the standard deviation of feature in each channel. By focusing on these aspects, function J can effectively guide our method in disrupting the generation of the neural style transfer.

Generation of perturbation. Algorithm 1 outlines the protection method transforming $\boldsymbol{x}$ to $\boldsymbol{x}^*$.[2] We employ an iterative approach with an ℓ_∞ norm constraint, denoted as $\|\boldsymbol{x}-\boldsymbol{x}^*\|_\infty \le \epsilon$. before the for loop, we randomly generate a small noise $\boldsymbol{\delta_0}$ to avoid the gradient is 0 in loops. For each iteration, the gradient of the loss function with respect to the input is computed, and the perturbations are updated in the direction maximizing the loss value, with a step size of α. The perturbations are then clipped to maintain the ℓ_∞ constraint.

3.3 Aesthetic Color Distance Metric

At present, there is no image distance metric specifically designed for distinguishing image differences affected by style from color perspective, i.e., the color consistency is often overlooked in existing research; the commonly used generic metrics are not well-suited for

[2] Our code is available at https://github.com/ZhongliangGuo/LAACA.

this task either, as evidenced by the experimental results summarized in Table 1. To address this limitation, we propose a new metric, Aesthetic Color Distance Metric (ACDM),[3] in the LAB color space, which is constructed to be perceptually uniform and more closely aligns with human visual perception compared to the RGB space [34]. The LAB color space consists of three channels: L represents lightness, ranging from 0 to 100, A and B represent color opponents, with A ranging from -128 (●) to 127 (●) and B ranging from -128 (●) to 127 (●).

Given two images $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^{C\times H\times W}$ in the LAB color space, where $C = 3$ denotes the number of color channels (L, A, B), and H, W denote the image height and width, respectively. For each color channel $c \in \{L, A, B\}$, we compute the histogram of values. The histogram represents the distribution of pixel values for each color channel, and is considered an effective method to characterize the color composition of an image.

The number of bins N_c for each channel is determined by taking the square root of the difference between the maximum and minimum values of that channel:

$$N_c = \left\lceil \sqrt{\max_c - \min_c} \right\rceil \tag{7}$$

where $\min_c$ and $\max_c$ denote the possible value range of channel c. Denote the bin width as $h = \frac{\max_c - \min_c}{N_c}$, the corresponding bins are $\{B_i\}_{i=1}^{N_c}$, where $B_i = [\min_c + (i-1)h, \min_c + ih)$ for $i = 1, \ldots, N_c - 1$, and $B_{N_c} = [\min_c + (N_c - 1)h, \max_c)$.

Let $\mathcal{H}_c : \mathbb{R}^{H\times W} \to \mathbb{R}^{N_c}$ denote the function that maps a channel image to its corresponding N_c-dimensional frequency count vector. To enable a meaningful comparison between the color distributions of the two images, we normalize the histogram vectors of both images $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ by dividing each histogram by its own sum:

$$\hat{\boldsymbol{h}}_c(\boldsymbol{z}_i) = \frac{\mathcal{H}_c(\boldsymbol{z}_{i,c})}{\sum_{j=1}^{N_c} \mathcal{H}_c(\boldsymbol{z}_{i,c})_j}. \tag{8}$$

This normalization step converts the histogram vectors into probability distributions, ensuring that the resulting normalized histograms $\hat{\boldsymbol{h}}_c(\boldsymbol{z}_1)$ and $\hat{\boldsymbol{h}}_c(\boldsymbol{z}_2)$ have a total sum equal to 1. By representing the color distributions as probability distributions, we can effectively capture the relative frequencies of pixel intensities in each channel, facilitating a fair comparison between the two images.

Earth Mover's Distance based metric. To measure the color difference between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, we employ the Earth Mover's Distance (EMD) between their normalized histograms for each channel. The EMD calculates the minimum cost required to transform one probability distribution into another. By using the EMD, we consider not only the absolute differences between corresponding histogram bins but also the overall shape and structure of the distributions. This is particularly important in the context of color distributions, as the EMD can capture perceptually meaningful differences that simple bin-wise comparisons may overlook. Moreover, the EMD takes into account the ground distance between bins, which allows for a more nuanced comparison of the color distributions.

When the two distributions are 1-D vectors, the distance can be solved with a closed-form solution [19, 22]:

$$\mathcal{D}_c(\boldsymbol{z}_1, \boldsymbol{z}_2) = \frac{\sum_{n=1}^{N_c} |\mathcal{F}(\hat{\boldsymbol{h}}_c(\boldsymbol{z}_1))_n - \mathcal{F}(\hat{\boldsymbol{h}}_c(\boldsymbol{z}_2))_n|}{N_c - 1}. \tag{9}$$

where $\mathcal{F}(\hat{\boldsymbol{h}}_c(\boldsymbol{z}_i))_n = \sum_{j=1\ldots n} \hat{\boldsymbol{h}}_c(\boldsymbol{z}_i)_j$ is the cumulative probability mass up to the n-th bin. To scale the data for better interpretability, we adopt a max-min normalization with a theoretical maximum difference of two distributions. By employing the EMD, we obtain a robust and semantically meaningful measure of the color difference between the two images in each perceptual color channel. Finally, we obtain the overall color difference score by summing the differences across all three channels:

$$\text{ACDM}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sum_{c\in\{L,A,B\}} \mathcal{D}_c(\boldsymbol{z}_1, \boldsymbol{z}_2). \tag{10}$$

Discussion on distance metrics. Consider a toy example of histograms from one channel of 3 images, each with 4 bins. Suppose the normalized histograms of these images are: $A = [0.0, 1.0, 0.0, 0.0]$, $B = [0.2, 0.3, 0.5, 0.0]$, $C = [0.2, 0.3, 0.0, 0.5]$. Histograms B and C have the same bin values but in a different order. Intuitively, the distribution of B ought to be more similar to A because the position of its 0.5 entry is closer to that of 1.0 in A compared with C. However, when we use ℓ_1 loss ($\mathcal{H}_1$), ℓ_2 loss ($\mathcal{H}_2$), Cross Entropy ($\mathcal{H}_3$), Cosine Similarity ($\mathcal{H}_4$) or Euclidean Distance ($\mathcal{H}_5$) to compare these vectors, we find that B and C have the same loss values when compared to A: $\mathcal{H}_1(A, B) = \mathcal{H}_1(A, C) = 0.35$, $\mathcal{H}_2(A, B) = \mathcal{H}_2(A, C) = 0.195$, $\mathcal{H}_3(A, B) = \mathcal{H}_3(A, C) = 1.4437$, $\mathcal{H}_4(A, B) = \mathcal{H}_4(A, C) = 0.4867$, $\mathcal{H}_5(A, B) = \mathcal{H}_5(A, C) = 0.8832$, which indicates that although B and C have different similarities to A, these three loss functions cannot distinguish between them. On the other hand, if we use EMD to compare these vectors, we find that: $\text{EMD}(A, B) = 0.7, \text{EMD}(A, C) = 1.2$.

EMD provides a more intuitive and accurate measure of the difference between the distributions. The smaller EMD value between A and B reflects that their high-value bins are concentrated in the same region, and only a slight movement of some pixels is needed to match them perfectly. In contrast, the larger EMD value between A and C indicates that more pixels need to be redistributed among the bins to match distributions.

4 Results

4.1 Experimental Setup

We use the normalized VGG-19 [10] as our encoder, consistent with NST methods like AdaIN and EFDM, capturing intermediate layer outputs from $relu1\_1, relu2\_1, relu3\_1, relu4\_1$. We set $k = 4$, $\alpha = 8$, $\epsilon = 80$, $T = 100$ to balance visual integrity and attack effectiveness, with hyperparameter discussions in the ablation studies. Content images are sourced from MS-COCO [23], and style images are from WikiArt [18].

We target five popular NST methods: Gatys [8], [4] AdaIN [15], [5] OST [26], [6] SANet [31], [7] and EFDM [42], [8] representing various approaches in the NST domain. For all NST methods, we set the image size as 512×512, for Gatys, we follow the initial setting, setting $style\_weight = 1e^6, content\_weight = 1, epochs = 500$; For OST, $\alpha = 0.6$; For AdaIN, SANet and EFDM, we apply the default parameter $\alpha = 1$ which was discussed in their paper. We randomly sample around 300 pairs of style and content images to evaluate original and protected artworks, pre-and post-protection NST images.

As no existing attack methods are specifically designed for style images of NST, we extend the Feature Disruptive Attack (FDA) [6]

[3] Our code is available at https://github.com/ZhongliangGuo/ACDM.

[4] https://pytorch.org/tutorials/advanced/neural_style_tutorial.html
[5] https://github.com/naoto0804/pytorch-AdaIN
[6] https://github.com/boomb0om/PyTorch-OptimalStyleTransfer
[7] https://github.com/GlebSBrykin/SANET
[8] https://github.com/YBZh/EFDM

to our task with the same layers and the encoder as our method. We also consider the Universal Adversarial Perturbation (UAP) [30] as a baseline due to its wide applicability in adversarial settings. To align with our method, the ℓ_ϵ norm of two baselines is also set as 80.

To further evaluate the performance of our method in real-world scenarios, where images are often compressed or downscaled for efficient distribution, we simulate common image degradation techniques by applying JPEG compression (retain 75% quality) and Gaussian blur (kernel in Equation 4 with $k = 3$) to the test images.

4.2 Evaluation for ACDM

Color is a key indicator for image style differences, often overlooked by existing metrics. Our proposed Color-based metric, ACDM, effectively distinguishes between style differences from color perspective.

Due to the current lack of well-annotated datasets for color-centric tasks, we validate ACDM in the Neural Style Transfer (NST) context. We hypothesize that positive pairs (same style, different content) should have higher color correlations than negative pairs (same content, different styles). We expect smaller ACDM scores for positive pairs and larger scores for negative pairs.

We sample 10,000 pairs from MS-COCO (content) and WikiArt (style), perform style transfer using EFDM, and compare ACDM with two popular metrics Structural Similarity Index Measure for color image (SSIMc) [39] and Learned Perceptual Image Patch Similarity (LPIPS) [41], where SSIMc is a SSIM's variant considering the color information. It is worth noting that we use the LPIPS with VGG, aligning with the convention in NST domain. The evaluation results are shown in Table 1.

Table 1. Evaluation on effectiveness of ACDM compared with SSIMc and LPIPS, ↑/↓ indicates that bigger/smaller value means better image quality.

IQA	positive pairs ($\mathbb{P}$)	negative pairs ($\mathbb{N}$)	change ratio $\frac{\lvert\mathbb{P}-\mathbb{N}\rvert}{\mathbb{P}}$
SSIMc ↑	0.2871	0.4072	41.83%
$\text{LPIPS}_{\text{VGG}}$ ↓	0.5851	0.5459	6.70%
ACDM ↓	0.0464	0.2982	**542.67%**

In comparison, although SSIMc considers the color information, it demonstrates a more notable ability to capture image structure, which is more related to the content of the image rather than its color style. This is evidenced by the higher SSIMc score for negative pairs, where the content is consistent, compared to positive pairs, where the content varies. This suggests that SSIMc is more sensitive to changes in image content rather than to changes in color or style. This perceptual ability of SSIMc can be leveraged as an evaluation metric for aspects other than color in our proposed Locally Adaptive Adversarial Color Attack (LAACA) method.

Furthermore, in the context of NST, LPIPS exhibits similar performance for both positive and negative pairs. This suggests that LPIPS may not be particularly sensitive to changes in either content or style when the other component remains consistent. In other words, LPIPS seems to be influenced by both content and style simultaneously, making it less discriminative when one of these factors is fixed. This observation highlights the need for a more targeted evaluation metric, such as our proposed ACDM, which can effectively capture color-related changes even when content or style is held constant.

These findings underscore the effectiveness of our proposed ACDM metric in evaluating color-mattered tasks. The substantial difference in scores between positive and negative pairs demonstrates its strong perceptual ability to capture color differences, setting it apart from other commonly used metrics.

4.3 Results for LAACA

We employ three evaluation methods to assess and analyze the experimental results: our proposed ACDM, SSIMc, LPIPS. ACDM quantifies the color variations between the original and attacked images. SSIMc is used to measure the changes in image structure before and after the attack, with higher SSIMc values indicating greater structural similarity and a maximum value of 1 indicating identical images. LPIPS measure the perceptual similarity of the pre- and post-attack images. For both SSIMc and ACDM, we use a Gaussian kernel size of 11, following the default setting of SSIMc, to ensure a consistent and comparable evaluation. In the following tables, $+/-$ indicates that bigger/smaller value means better results.

SSIMc. Table 2 presents the SSIMc scores comparing the structural similarity between the original and attacked images. Our method overwhelmingly achieves an SSIMc score better than that of FDA and UAP, indicating excellent preservation of structural information. For the style-transferred images, LAACA obtains an average SSIMc score of 0.3356, better than both FDA and UAP, demonstrating its effectiveness in disrupting the style transfer process. Under defense, LAACA's performance slightly declines. However, these results remain within an acceptable range, showcasing LAACA's robustness in real-world scenarios where images may undergo compression or blurring during distribution.

Table 2. SSIMc evaluation results, a higher score means better quality.

(**the best**, <u>the second best</u>)	style images $^+$	Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average $^-$
LAACA	<u>0.6130</u>	0.2392	0.3891	0.3671	0.3150	0.3674	**0.3356**
FDA	0.5647	0.3075	0.5260	0.4503	0.4016	0.5061	0.4593
UAP	0.3556	0.3059	0.4555	0.3121	0.2599	0.4222	<u>0.3511</u>
JPEG 75% Comp.	**0.6218**	0.2639	0.3986	0.3757	0.3229	0.3765	0.3475
Gaussian blur	**0.6214**	0.5499	0.5354	0.5403	0.4704	0.5080	0.5208

ACDM. Table 3 presents the ACDM scores, which measure the color difference between the original and attacked images. For the style images, our proposed LAACA method achieves an ACDM score of 0.0495, significantly better than FDA (0.1406) and UAP (0.1350), indicating excellent preservation of visual color information in protected images. For the style-transferred images, LAACA obtains the second highest average ACDM score, demonstrating its ability to disrupt the color of the style-transferred images. Despite the application of common image distortions, such as JPEG compression and Gaussian blur, LAACA maintains its effectiveness in both preserving color information in the style images and disrupting color in the style-transferred images, as evidenced by the consistent ACDM scores across all scenarios.

LPIPS. The LPIPS results in Table 4 demonstrate that LAACA ranks in the top tier for original/protected style images in terms of maintaining the perception, and it performs the best on disrupting the perception of post-NST images. This indicates that LAACA is effective in preserving the perceptual quality of the style images while successfully disrupting the perceptual similarity of the NST images. When subjected to defense methods, LAACA shows competitive results. JPEG compression performs similarly to LAACA on both style images and NST images, suggesting that LAACA maintains its effectiveness even when the images undergo compression. Although Gaussian blur slightly degrades LAACA's performance, the performance loss remains within an acceptable range. This indicates that our method exhibits a certain level of robustness against potential image degradation that may occur during distribution.

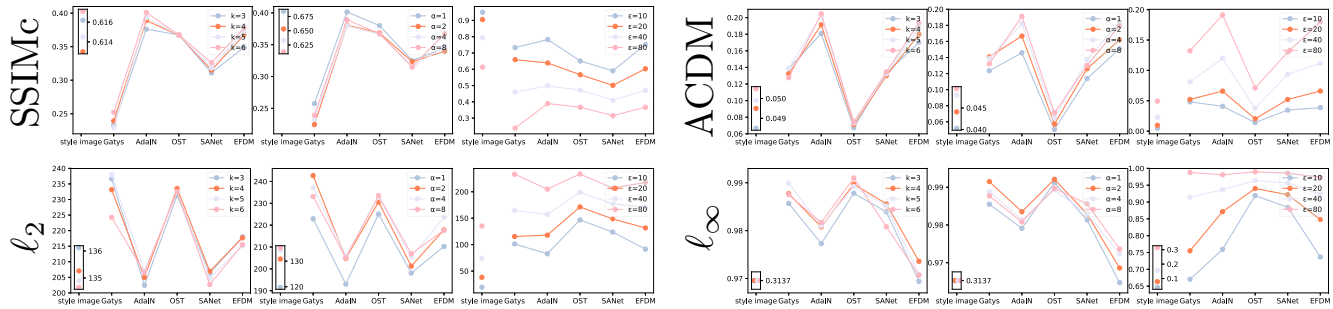

Figure 4. Ablation study results, for those labels on X-axis, they represent comparisons of pre- and post-attack images of X-axis value on a certain metric.

Table 3. ACDM evaluation results, a lower score means better quality.

(**the best**, the second best)	style images$^-$	Neural Style Transfer Methods					Average $^+$
		Gatys	AdaIN	OST	SANet	EFDM	
LAACA	0.0495	0.1322	0.1913	0.0711	0.1300	0.1798	0.1409
FDA	0.1406	0.1746	0.1553	0.1218	0.1517	0.1473	**0.1486**
UAP	0.1350	0.1475	0.1412	0.1208	0.1406	0.1317	0.1364
JPEG 75% Comp.	0.0492	0.1244	0.1821	0.0682	0.1241	0.1694	0.1196
Gaussian blur	**0.0438**	0.0660	0.0839	0.0493	0.0906	0.0875	0.0755

Table 4. LPIPS evaluation results, a lower score means better quality.

(**the best**, the second best)	style images$^-$	Neural Style Transfer Methods					Average$^+$
		Gatys	AdaIN	OST	SANet	EFDM	
LAACA	**0.4043**	0.6080	0.5552	0.4949	0.5545	0.5723	**0.5570**
FDA	0.4835	0.5841	0.4476	0.4437	0.5151	0.4642	0.4897
UAP	0.5740	0.5550	0.4239	0.4850	0.5231	0.4425	0.4859
JPEG 75% Comp.	**0.3985**	0.5944	0.5487	0.4881	0.5492	0.5653	0.5491
Gaussian blur	0.4226	0.4503	0.4513	0.4129	0.4615	0.4647	0.4481

4.4 Ablation Studies

In ablation studies, in addition to SSIMc and ACDM, we also include the L_p norms as evaluation metrics to quantify the pixel-level differences between the protected and original images, which is regarded as a convention in adversarial attack researches. For all experiments, we set $T = 100$, ensuring convergence.

The parameter k determines the separation of high-frequency and low-frequency components in the image, with a larger k resulting in a wider range of frequencies being considered as high-frequency. As k increases from 3 to 6, the SSIMc scores for the style images show a slight improvement, indicating that a larger k may lead to a better separation of high-frequency and low-frequency regions, resulting in improved structural similarity between the pre- and post-attack style images. However, the SSIMc scores for the other NST methods decrease, suggesting that a larger k may compromise their structural similarity. The ACDM scores increase with larger k, indicating that a larger k may introduce more color distortions in the style images while affecting the color stability of other NST methods. The ℓ_2 distance decreases slightly for the style images as k increases, implying that a larger k may produce lower pixel-wise differences, but it increases for the other NST methods. The ℓ_∞ scores remain relatively stable across different k values for all methods, indicating that the maximum pixel-wise difference is nearly unrelated to k.

The parameter α determines the magnitude of the update in each attack iteration. As α increases from 1 to 8, the SSIMc scores for the style images remain relatively stable, while the scores for the other NST methods decrease, suggesting that larger step sizes may compromise their structural similarity. The ACDM scores increase with larger α values, indicating that a larger step size may lead to more color distortions in the style images while also affecting the color stability of other NST methods. The ℓ_2 distance increases for all methods with higher α values, implying that larger step sizes may result in greater pixel-wise differences between the pre- and post-attack images. The ℓ_∞ scores remain relatively stable across different α values for all methods, suggesting that the maximum pixel-wise difference is not significantly influenced by α.

The parameter ϵ defines the maximum allowed deviation from the original image in the style transfer process. As ϵ increases from 10 to 80, the SSIMc scores for the style images remain relatively high, while the scores for the other NST methods decrease significantly, indicating that a larger perturbation range leads to a substantial decrease in their structural similarity. The ACDM scores increase with larger ϵ values, suggesting that a higher perturbation range introduces more color distortions in the style images while also affecting the color stability of other NST methods. The ℓ_2 distance increases for all methods as ϵ grows, implying that a larger perturbation range results in greater pixel-wise differences between the pre- and post-attack images. The ℓ_∞ scores increase with higher ϵ values, indicating that the maximum pixel-wise difference between pre- and post-attack images becomes larger as ϵ increases, particularly for NSTs.

5 Conclusion, Limitations, and Future Work

In this work, we propose the Locally Adaptive Adversarial Color Attack, a method designed to interrupt unauthorized neural style transfer use cases. Our approach significantly degrades the quality of NST outputs while introducing acceptable perturbations, which will discourage potential infringers from using the protected artwork, because of the bad NST generation. To supplement metrics in evaluating the performance of color-mattered tasks, we introduce an IQA, ACDM, which quantifies the color distortions between pre- and post-attack style images. The evaluation of ACDM's performance in experiments validates its effectiveness in assessing color-related tasks. Experiments demonstrate the efficacy of our attack method in compromising the style transfer process, resulting in significant color distortions and structural dissimilarities in NST images while maintaining the acceptable visual integrity of the post-attack style images. However, our work has some limitations. The method's runtime on common GPUs is not optimal, and the numerous hyperparameters currently provided may not be suitable for all images. Categorizing style images into abstract and realistic paintings, our method is more effective on abstract paintings, possibly because their bolder colors and high-frequency components provide a larger manipulation space. Future work will focus on improving computational efficiency and exploring ways to make these parameters adaptive. Overall, our approach offers artists a potential tool to protect their intellectual property, with the promise of mitigating or curbing electronic IP infringement.

References

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 248–255. IEEE Computer Society, 2009.

[3] J. Dong and X. Xie. Visually maintained image disturbance against deepfake face swapping. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[4] J. Dong, Y. Wang, J. Lai, and X. Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 2023.

[5] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] A. Ganeshan, V. B.S., and R. V. Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[7] M. Garg, J. S. Ubhi, and A. K. Aggarwal. Neural style transfer for image steganography and destylization with supervised image to image translation. *Multimedia Tools and Applications*, 82(4):6271–6288, 2023.

[8] L. Gatys, A. Ecker, and M. Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326–326, 2016.

[9] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[12] C. Guo, J. S. Frank, and K. Q. Weinberger. Low frequency adversarial perturbation. In *Uncertainty in Artificial Intelligence*, pages 1127–1137. PMLR, 2020.

[13] Z. Guo, W. Li, Y. Qian, O. Arandjelović, and L. Fang. A white-box false positive adversarial attack method on contrastive loss-based offline handwritten signature verification models. *arXiv preprint arXiv:2308.08925*, 2023.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[15] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang. Exploring frequency adversarial attacks for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4103–4112, 2022.

[17] N. Kalischek, J. D. Wegner, and K. Schindler. In the light of feature distributions: Moment matching for neural style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9382–9391, 2021.

[18] N. Kiri and W. Kan. Painter by numbers, 2016. URL https://kaggle.com/competitions/painter-by-numbers.

[19] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

[20] G. E. Legge and J. M. Foley. Contrast masking in human vision. *Josa*, 70(12):1458–1471, 1980.

[21] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems*, 30, 2017.

[22] H. Lin, Z. Ma, R. Ji, Y. Wang, Z. Su, X. Hong, and D. Meng. Semi-supervised counting via pixel-by-pixel density distribution modelling. *arXiv preprint arXiv:2402.15297*, 2024.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[24] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14973–14982, June 2022.

[25] M. S. Livingstone. Art, illusion and the visual system. *Scientific American*, 258(1):78–85, 1988.

[26] M. Lu, H. Zhao, A. Yao, Y. Chen, F. Xu, and L. Zhang. A closed-form solution to universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[29] S. R. Maiya, M. Ehrlich, V. Agarwal, S.-N. Lim, T. Goldstein, and A. Shrivastava. Unifying the harmonic analysis of adversarial attacks and robustness. In *BMVC*, 2023.

[30] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] D. Y. Park and K. H. Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] E. Risser, P. Wilmot, and C. Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017.

[34] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao. Statistics of cone responses to natural images: implications for visual coding. *JOSA A*, 15 (8):2036–2045, 1998.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. Computational and Biological Learning Society, 2015.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[37] C. Vallance. Ai image creator faces uk and us legal challenges. https://www.bbc.co.uk/news/technology-64285227, 2023.

[38] K. Wang, J. Shi, and W. Wang. Lfaa: Crafting transferable targeted adversarial examples with low-frequency perturbations. In *ECAI 2023*, pages 2483–2490. IOS Press, 2023.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[40] S. Zeki. *Inner vision: An exploration of art and the brain*. Oxford University Press, USA, 2000.

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[42] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8035–8045, 2022.

[43] S. Zhao, J. Wen, A. Luu, J. Zhao, and J. Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12303–12317, 2023.

[44] S. Zhao, L. Gan, L. A. Tuan, J. Fu, L. Lyu, M. Jia, and J. Wen. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.12168*, 2024.

[45] S. Zhao, M. Jia, Z. Guo, L. Gan, J. Fu, Y. Feng, F. Pan, and L. A. Tuan. A survey of backdoor attacks and defenses on large language models: Implications for security measures. *arXiv preprint arXiv:2406.06852*, 2024.

[46] S. Zhao, M. Jia, L. A. Tuan, F. Pan, and J. Wen. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. *arXiv preprint arXiv:2401.05949*, 2024.

[47] S. Zhao, L. A. Tuan, J. Fu, J. Wen, and W. Luo. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–11, 2024. doi: 10.1109/TASLP.2024.3407571.